Information Engineering for Molecular Diagnostics

James M. Sorace , M.D., Michele Ritondo, M.S., Kip Canfield Ph.D.
Department of Pathology and Laboratory Medicine, Baltimore VAMC,
Department of Pathology at The University of Maryland and
The Laboratory for Health Care Informatics at The University of Maryland Baltimore County

ABSTRACT

Clinical laboratories are beginning to apply the recent advances in molecular biology to the testing of patient samples. The emerging field of Molecular Diagnostics will require a new Molecular Diagnostics Laboratory Information System which handles the data types, samples and test methods found in this field. The system must be very flexible in regards to supporting ad-hoc queries. The requirements which are shaping the developments in this field are reviewed and a data model developed. Several queries which demonstrate the data models ability to support the information needs of this area have been developed and run. These results demonstrate the ability of the purposed data model to meet the current and projected needs of this rapidly expanding field.

INTRODUCTION

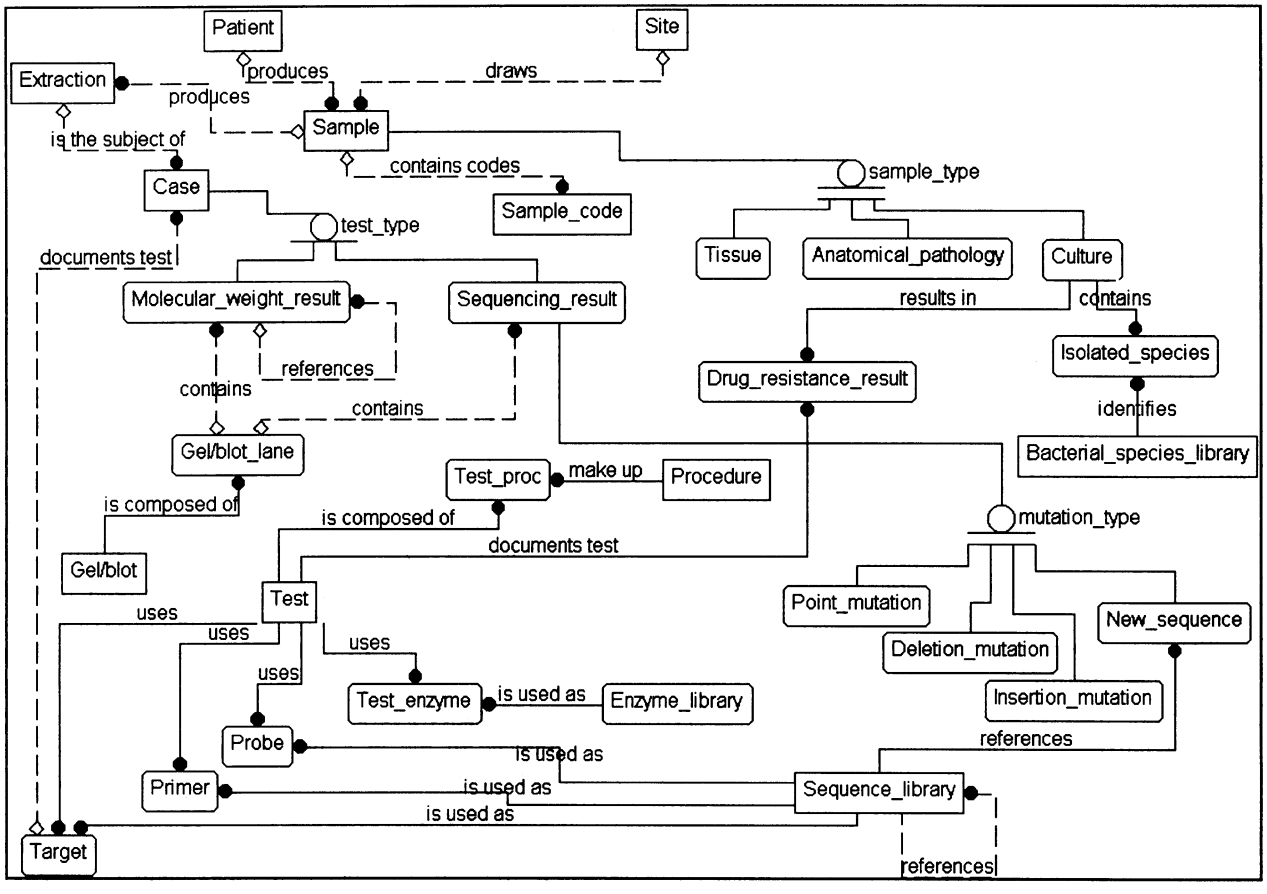
Molecular Diagnostics (MD) is a new family of techniques which will greatly increase the type and amount of information generated on clinical samples. MD is currently enjoying considerable success in two major areas. The first is the area of Clinical Microbiology, which involves the detection and characterization of infectious agents [1]. Both the diagnosis for therapeutic purposes, and characterization of the pathogen for epidemiological purposes are important. Interest in this data is greatly increasing due to the rapid rise of antibiotic resistance and the spread of new diseases. The second area is in the field of Anatomic Pathology, in which molecular techniques are being used to characterize infectious agents directly from tissue samples, the molecular changes underlying malignancy, and the determination of malignant cellular lineage [2].

Despite intense interest in informatics support for molecular biology, none of the current databases are principally designed to support clinical research in this field [3,4]. In order to address this issue we are developing a data model for a Molecular Diagnostic Laboratory Information System (MDLIS). The database resulting from this model allows for rapid searching and retrieval of information for clinical, research, and epidemiological needs. While numerous databases and software systems exist for biological sequence comparisons, key elements for an MDLIS are lacking. These elements fall into three main areas. First there is the need to actually manage and store the work flow of the laboratory. This includes keeping track of tests ordered, inventorying DNA extractions, and archiving test result data. Secondly, the system needs to be able to support a very broad range of potential queries. For example, the species of the bacterial isolate and its antibiotic resistance, the microscopic diagnosis and morphology/topography codes of a tumor, or the sequence of the PCR primers used to generate the result, may all be relevant information. Thirdly, there is a large communication need in this area. Laboratories will need to share data, methods, and new discoveries regarding the genetic changes regarding diseases states.

REQUIREMENTS ANALYSIS

MD data has several unique requirements in database design. For example, several entities not typically found in hospital laboratory information systems must be stored and manipulated. These include sequence data, and molecular weight data (MWD). In MWD molecules are separated by size using gel electrophoresis (the presence or absence of a specific band is detected), and the molecular weight of the resulting band is then calculated. In addition, MWD is unique in that different interpretations of the same data type can be generated by differing assays. For example, both a restriction enzyme digest of a bacterial isolate and the PCR amplification of a bacterial gene produce MWD. However in the first case, the test will produce multiple bands while in the second case the test will produce a single band. In some cases, tests can be multiplexed so that a single assay can produce multiple molecular weight bands, each representing a different target molecule. Besides MWD, sequence data can be generated. A sequence mutation can involve point mutations, deletions, or insertions.

0195-4210/94/$5.00



igure 1: Data model for Molecular Diagnostics.

MD also has a more complex specimen work flow than most other areas of the clinical laboratory [5]. Nucleic acid (DNA or RNA) must first be extracted from the sample. These extracts are then inventoried and serve as the starting material for additional tests. The sources of the samples themselves are very diverse. Tests may be run on specimens received directly from patients (blood, amniotic fluid etc.), from surgical pathology samples which have been processed for histological examination, or from bacterial cultures which have already been isolated by the microbiology laboratory. Thus the nucleic acid extract must be linked to a detailed record of the sample's history. The nature of the target sequence present in the extract is also broad. Tests may be run to detect alleles present in the normal population, the presence of mutations and translocations found in a tumor, the pulse field electrophoresis pattern of an entire bacterial genome, or the presence of an antibiotic resistance plasmid in a microbiology isolate.

MD will also have very significant communication requirements. Laboratories will need to share data , methods, and new discoveries regarding the mutations found in disease states. These issues will only increase as data from the human genome project becomes available and the number of genetic test increases. Also current goals of the human genome project include increasing the rate of DNA sequencing methods. When these methods are applied to clinical microbiology samples with the inherent genomic variability of bacterial and viral isolates, the number of comparisons which need to be made between laboratories will increase. This type of communication can best be insured if the databases use a common data model or a subset of a common data model [6]. Such a data model can also form the basis of specialized repository databases for the molecular diagnostics community. It is also important to supply adequate links to currently available databases like GenBank, so that the database can be readily cross-referenced.

METHODS

In order to begin to develop software which will meet

the needs as outlined above, our group has performed a detailed requirements analysis, including the fields of anatomic pathology and clinical microbiology. Several iterations of model building (including data entry trials) resulted in the data model displayed in figure 1. This entity-level model is relational. It can generate SQL to create the relational database in that each entity is a table, there are no many-to-many relationships, and there are no ternary (or higher) relationships (a complete text version of the model is available by anonymous file ftp from 130.85.105.8). The model has several features relevant to the issues discussed, and outlines a rational guide towards future growth.

First, patient samples can be subgrouped as either tissue, anatomic pathology, or clinical microbiology samples in a one-patient-to-many-samples relationship. Entities covering these possibilities contain information unique to each specimen source. In the case of culture (clinical microbiology) samples, additional tables are present which allow the entry of antibiotic resistance and strain-isolated data. Thus it is possible to retrieve data based on queries searching for this important information. This structure also clearly separates results which could otherwise be misinterpreted. A tissue biopsy with appropriate PCR testing may reveal a mixed pattern of infection , while the microbiology sample may only contain a single isolate which had a growth advantage under the laboratory culture conditions. Anatomic pathology samples can be searched for under morphology/topography codes or ICD-9 codes, and by tracking the sample's surgical pathology number, other forms of information (microscopic images of tissue samples, links to the patients medical record) found in the hospital's information system can be retrieved.

We have developed a Test library, and a Sequence library which acts as the data dictionary for the data model. Each test can be linked to the target, probe, primer and enzyme tables as appropriate. A test is given a unique Test_ID and it is linked in a one-to-many relationship with a procedure library. The procedure entity provides a field for the classification of the procedure (e.g., extraction, pulse field electrophoresis etc.), and a memo field describing the method. This allows each test to be defined as an ordered set of procedures and assures the laboratory staff a modular procedure library.

Perhaps the most challenging aspect of MD informatics lies in the archiving and interpretation of the laboratory results. The Case entity links the nucleic acid sample being tested to the test being performed and the test target sequence (thus supporting multiplexed assays with more than one target sequence). Results of each unique case can then be recovered under the Molecular_weight_result or Sequence_result entities. These sections of the data model contains sequence, and MWD data types. Support is also available for the entry of text results were appropriate. Given the complexities of the MWD, the database designer faces the prospect of numerous entities optimized for each potential combination of data, or a single entity which is more comprehensive. The model described here implements the latter choice. Each band in the Molecular_weight_result table has its molecular weight, and intensity recorded, is assigned to the gel and lane in which it was detected and linked via its Case_ID to the Test_ID and Sample_ID used to generate it. A case can have many such bands related to it, and the bands can be in one lane of one gel or widely dispersed over many gel/lane combinations.. Gel images can also be stored and recalled when appropriate.

The Case entity also links to the Sequence_result entity. Sequence mutations are then stored in specific entities for point mutations, deletions, insertions, or new sequences. The tables support the entry of mutations in a variety of formats including the codon which is involved and the nature (substitution, frame shift etc.) of the change. Again the results can also be linked to the Sequence_library, allowing the target sequence to be quickly recovered for comparison.

The Sequence_library contains several types of information. First, it contains the actual sequences of the targets, probes, and primers specified in the test library, and it can also store new sequences generated by laboratory testing. By including entries for the GenBank number, and the Genome Database number, other databases can be quickly cross-referenced. Other relevant information such as the bacterial species/target protein and genomic locus, can be stored. This table is heavily cross-referenced throughout the database with links to the target, primer, and probe tables. The actual sequence of the nucleic acid can be stored in this table, enabling rapid access to state-of-the-art sequence search and comparison algorithms.

TABLE 1

Output for query 1; see text for details.

GEL NUMBER	LANE NUMBER	SEQUENCE #	TEST ID	SAMPLE ID	CASE ID
1	6	1	1	3172	5
1	7	1	1	3774	6
2	2	1	1	3530	8

TABLE 2

Output for query 2; see text for details.

SEQUENCE #	CASE ID	CODON	FROM	TO	EXON	A.A. FROM	A.A. TO
2	WXYZ100	286	GAA	AAA	8	Glu	Lys
2	WXYZ110	273	CGT	CAT	8	Arg	His
2	WXYZ131	282	CGG	GGG	8	Arg	Gly
2	WXYZ135	273	CGT	CAT	8	Arg	His
2	WXYZ140	269	AGC	CGC	8	Ser	Arg
2	WXYZ600	271	GAG	AAG	8	Glu	Lys
2	WXYZ80	282	CGG	TGG	8	Arg	Trp
2	WXYZ356	273	CGT	CAT	8	Arg	His
2	WXYZ22	280	AGA	AAA	8	Arg	Lys
2	WXYZ177	282	CGG	TGG	8	Arg	Trp
2	WXYZ484	275	TGT	TCT	8	Cys	Ser

TABLE 3

Output for query 3; see text for details.

SEQUENCE #	CASE ID	EXON	CODON DELETED
2	WXYZ11	5	157
2	WXYZ12	5	158
2	WXYZ14	5	159

QUERY SUPPORT

Given the data model as outlined above, it is important to demonstrate its ability to support a wide range of scientifically and clinically relevant ad hoc queries. This is especially true given the extensive communications needs which will develop in this field. We performed a trial using a Microsoft ACCESS implementation of the data model. This is currently implemented on a stand-alone PC. Data was entered from several sources. First, Hepatitis C PCR test result data, bacterial restriction fragment length polymorphism, and antibiotic resistance data were entered from the Molecular Diagnostics Laboratory at the Baltimore VA Medical Center (MDVA) . Point mutation data for P53 oncogene mutations from Breast Cancers obtained from the Armed Forces Institute of Pathology (AFIP) were also entered. Based on the current and projected needs of the field we developed 20 queries which include:

(1) Be able to assist in epidemiological research by searching for restriction enzyme patterns of bacterial pathogens. This allows for comparisons of species isolated between laboratories. An example of this type of query is shown in Table 1, This query searches the database for all samples which have a specified number of bands within a defined molecular weight range for a given test and target sequence. In this instance the database has been searched for all Hha1 digest (Test_ID =1) of all Streptococcus Agalactae gb B genomes (Sequence_number = 1) which have 3 bands present in the 6,000-to-15,000 molecular weight range. In this query the Sample, Extraction, and Molecular-weight-result entities were linked

(2) Retrieve all the samples with a given point mutation in a given codon range in a given sequence. Table 2 illustrates the output of such a query searching for point mutations between codons 250 and 290 of the p53 oncogene (Sequence_number=2). The nucleic acid sequence of the normal and mutated codon and the resulting

amino acid change are also presented, as is the Case_ID and the involved exon.
(3) Summarize all the known deletion mutations of a given sequence. This was done for the P53 ocogene. Table 3 shows the output of such a query , which presents the involved exon, and the deleted codon.

Both queries 2 and 3 involved linking the Case and Sequencing_result entities with either the Point-mutation or Deletion_mutation entites.

DISCUSSION AND CONCLUSIONS

MD will significantly change both the practice of medicine and the information system requirements for the clinical laboratory. We have presented a data model for discussion that meets several of these requirements. These requirements can be grouped into 2 basic categories: communication and content.

First, a standard data model after refinement and user trials allows an infrastructure to support communications in the rapidly growing field of MD. The common fields and structure of the model simplify data sharing and database querying. Data sharing is made possible because the content and structure of the database model standard is understood by all users. Furthermore, simple database scripts allow users to import and export data over a network. This feature allows the development of intelligent software agents to support and manage the creation and update of research repositories that receive data from many sites. We are now implementing a trial of this technology for a research repository.

Query from multiple sites is also simplified by a standard data model. Not only the data but also supporting data dictionaries can be transferred over networks using the same database script methodology. These scripts are text files containing the SQL statements necessary to perform the desired actions on the target database. The SQL statements can be wrapped in a language/protocol such as Knowledge Query and Manipulation Language (KQML) [7] to supply the transport and per formatives for use on an network using intellegent agents. The creation and sharing of useful user interfaces is also simplified with a standard data model as the programmer is freed from worrying about data structures and compatibility. Querying of repositories could accomplish such tasks such as the geographic distribution of mutations. This type of query has already proven important as a P53 codon 249 mutation in hepatocellulare carcinoma is associated with environmental exposure to aflatoxin B1 [8]. The ability to search for mutation distribution based on patient zipcode might aid in this type of query. Without a common data model these and other types of queries are hindered.

Finally, the content of this data model explicitly defines our vision of what the needed requirements are for this field. This allows others in the field to respond to and improve this model with additions and deletions. Such a dialog is essential to standards initiation and only possible in a published forum.

ACKNOWLEDGMENTS

The authors thank Jack H. Lichy, M.D., Ph.D., and Jeffery K. Taubenberger, M.D., Ph.D., of the AFIP and Judy J. Johnson, Ph.D. of the BVAMC for sharing their data. We also thank Lawrence Brown, M.D. for help in preparing this manuscript. This material is based in part upon work supported under a National Science Foundation Graduate Research Fellowship.

REFERENCES

[1] J. Veralovic et. al, DNA-Based identification and epidemologic typing of bacterial pathogens. Archives of Pathology and Laboratory Medicine 117: 1088-1090, 1993.
[2] J. Rowely et. al, The clinical application of new DNA diagnostic technology on the management of cancer patients. JAMA 270: 2331-2337, 1993.
[3] H. Bilosky, C. Burks, The GenBank genetic sequence data bank. Nucleic Acid Research ,16: 1861-1863, 1988.
[4] P. Pearson, The GDB human genome database. Nucleic Acid Research 20: sup 2201-2206, 1992.
[5] B. McCreedy, T. Calloway, Laboratory Design and Work Flow, in Diagnostic Molecular Biology: Principle and Applications, D. Persing et. al eds., American Society for Microbiology. Washington DC, 1993.
[6] K. Canfield et. al, The standard data model approach to patient record transfer, TR Ihi-ifsm-umbc-jan-1994.
.[7] Finin T., et al., Specification of the KQML Agent-Communication Language. Technical report EIT TR 92-04, Enterprise Integration Technologies, Palo Alto, Ca, 1992.
[8] C. Harris, p53: At the crossroads of molecular carcinogenesis and risk assessment. Science 262:1980-1981, 1994.